University of California, Berkeley

Final Project of DATA-8

Summer 2020

Airbnb Market under COVID-19 in San Francisco

Hypothesis Testing with Airbnb Availability Rate Using Python

Yun Lin

collin61514@gmail.com

16 August 2020

Table of Content

Abstract

Airbnb offers a platform to connect hosts with guests for short-term or long-term lodging accommodations. Compared to similar firms offering vacation rental services such as VRBO or HomeAway, Airbnb is the largest and most prominent, with more than 7 million listings worldwide and 2 million people staying in one of its listings per night in 2018. Since its founding in 2008, hosts on the platform have served more than 750 million guests, and the firm has grown at an exponential rate globally pre-COVID.

The data presented are completely from web scraping the Airbnb website in June 2020 for random subset of listings in San Francisco. As a result, the data only contain information that a visitor to Airbnb's site can see. This includes the listings table that records all Airbnb units and the calendar table that records the availabilities for the next 365 days and quoted price per night over the next year of each listing. What each table specifically describes will be gone over in the Data Description section below. Note that we do not observe Airbnb transactions or bookings, but only the dates that are available or unavailable through calendar.

Data Description

The dataset consists of many tables stored in the data folder, but in this research, we will only be using calendar.

The dataset of calendar contains each listing's availability and price over the next year. This data is the same as the calendar that pops up when users try to select the dates of a reservation for a particular listing. For example, the first row means that the listing with ID 40138 was not available on June 8th, 2020. The price per night of this listing is $67.

calendar:

- listing_id: ID of airbnb listing
- date: date of the potential availability in question
- price: price per night of listing in USD
- available: true or false value representing whether the listing was available.

Introduction

In this project, we are exploring the data of Inside Airbnb, an independent project that collects and hosts Airbnb data on 100 or more cities around the world. Thankfully, Inside Airbnb made their data public so we can explore their data freely. What we are using is the Airbnb data from San Francisco collected in June 2020 and 2019. We are comparing two datasets in different year in order to see how much COVID-19 affected Airbnb in San Francisco. From the given datasets, we will be using 'calendar_2019' and 'calender_2020'. Looking at our dataset 'calendar', this dataset consists of features, 'listing_id', 'date', 'price', and 'available'. The 'listing_id' is used to distinguish the data points, 'date' gives the potential availability in question, 'price' gives price per night of listing in USD, and finally 'available' gives true or false value representing whether the listing was available. After reviewing our datasets' features, we decided to focus on 'listing_id' and 'available' in order to check the availability before after COVID-19 occured. These variables will be the most important feature to our analysis since it directly shows whether availability changed between two years by true and false. We can count the number of true and false, draw a distribution chart, and get percentage chart in order to see the analysis this question in multiple views

Hypothesis Testing and Prediction Questions

Our team (Yun and Andy) were both affected by COVID-19 earlier this year on shelter problems. We both had experience of searching on Airbnb to find a place for 2 weeks of self-isolation in our home country. This experience motivated us to analyze on how the Airbnb's availability changed by COVID-19. Specifically, we are focusing on how the availability changed on year 2020, the year COVID-19 started, and before year 2020. We suspected that the percentage of Airbnb availability days went up after COVID-19 since there were less travelers. Our null hypothesis is the distributions of the available days for Airbnb did change after the COVID-19.

A/B Testing

Since we want to compare values from two year, 2019 and 2020, it is best to use A/B testing for our hypothesis testing. For the prediction part, our team is using the price mean of 2019 and availability of 2019 to create the regression line to predict the availability for 2020.

Exploratory Data Analysis

This table shows the difference of available days for same place before and after COVID-19.

The features of this table consists of:
listing_id: ID of airbnb listing
availability before COVID-19: the number of available days listed on the year 2019
availability after COVID-19: the number of available days listed on the year 2020
change: The difference between number of available days between the year 2019 and 2020.
price mean: The average of Airbnb price over the year

The table below is significant since it shows trend of how the available days has increased after the COVID-19. Therefore, we suspected that the available days has increased due to COVID-19 suspecting due to lack of international travelers.

listing_id	availability before Covid-19	availability after Covid-19	Availability change
5858	336	365	29
8339	90	365	275
10820	306	364	58
12041	365	365	0
12584	269	365	96

... (404 rows omitted)

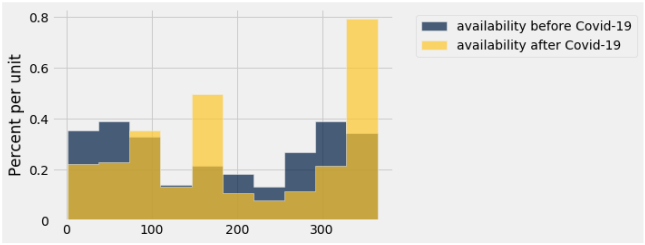
The table below shows the percentage difference of available days for same place before and after COVID-19. The features of this table consists of:
listing_id: ID of airbnb listing
% availability before COVID-19: the percentage of available days on the year 2019
% availability after COVID-19: the percentage of available days on the year 2020
change: The difference between percentage of available days between the year 2019 and 2020.

id	% availability before Covid-19	% availability after Covid-19	change
5858	92.0548	100	7.94521
8339	24.6575	100	75.3425
10820	83.8356	99.726	15.8904
12041	100	100	0
12584	73.6986	100	26.3014

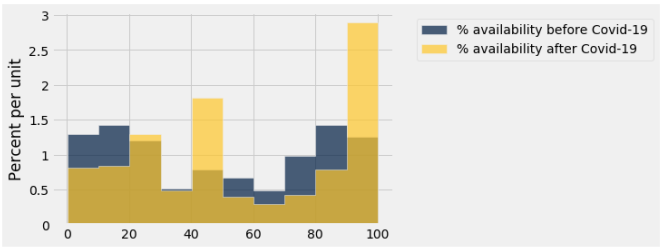
... (404 rows omitted)

Avability of Airbnb before and after COVID-19 Comparsion graph

```
In [123]: combine.hist('availability before Covid-19','availability after Covid-19')
```




```
In [124]: per_combine.hist('% availability before Covid-19','% availability after Covid-19')
```



availability before Covid-19:the number of available days listed on the year 2019
availability after Covid-19:the number of available days listed on the year 2020

The histogram shows the availability before and after Covid-19.

Null hypothesis

Looking at our dataset, we suspected there was an increase in days available for places in Airbnb. We wanted to check if this prediction is true or not which led to this analysis. **Our null hypothesis is the distributions of the available days for Airbnb did change after the COVID-19.** Our alternative hypothesis is that the distributions of the available days for Airbnb did not actually change after the COVID-19 which means COVID-19 impacted Airbnb's job. We decide to check this with a 0.05 level of significance. Since we want to compare values from two year, 2019 and 2020, it is best to use A/B testing for our hypothesis testing.

Availability difference Simulation distribution

We build up the distribution model with 1000 simulation.

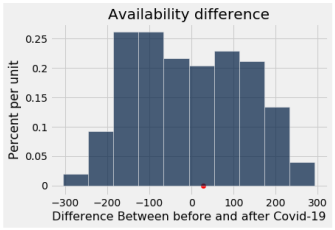
Looking at our observed difference, it is included in our 95% confidence interval. This means that our observed difference from the dataset exploration seems reasonable which leads to rejecting our null hypothesis.

Therefore, we **reject our null hypothesis** and approve our alternative hypothesis. The distributions of the available days in Airbnb did not change after COVID-19.

```
ava_sample= make_array()
repetitions =1000
for i in np.arange(repetitions):
    a=one_simulated_difference(combine,'availability before Covid-19','availability after Covid-19')
    ava_samples= np.append(samples,a)


Table().with_column('Difference Between before and after Covid-19', ava_samples).hist()
plots.title('Availability difference')
plots.scatter(observed_difference, 0, color='red', s=40)
```

```
<matplotlib.collections.PathCollection at 0x7ff80e8a2e80>
```




```
a=percentile(2.5,ava_samples)
b=percentile(97.5,ava_samples)
print('95% confidence interval range is [',a,',',b,']')
```

```
95% confidence interval range is [ -227.25 , 233.0 ]
```

Calculation

We calculated the correlation, slope, intercept and prediction formula respectively and printed.

Correlation between the airbnb price and availability is -0.1158

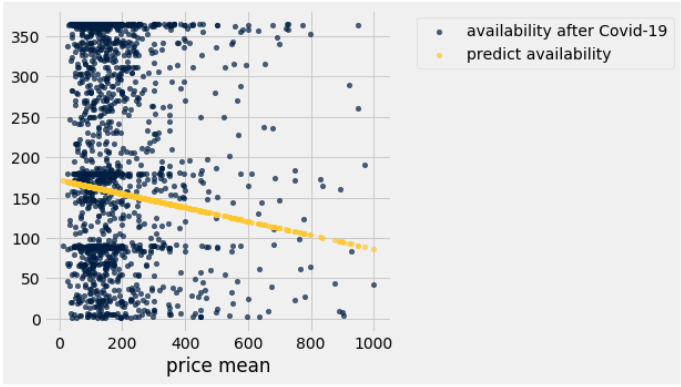
Slope between the airbnb price and availability between the airbnb price and availability is -0.08.

intercept between the Airbnb price and availability is 171.85

Perdition availability after Covid-19 = -0.08 * price before Covid-19 + 171.85

Regression line

```
after_plot=pred_after.select('availability after Covid-19','price mean','predict availability')
after_plot.scatter('price mean')
```



Data flaws

How does the scatter plot not it fit in regression line

For the prediction part, our team is using the price mean of 2019 and availability of 2019 to create the regression line to predict the availability for 2020. There are three reasons why the data does not fit in the regression line,

1. Not all the Airbnb is open for 365 which their maximum available day is limited, Availability is showing whether the Airbnb is booked or not, but when the customer choosing the place, they will normally look in to the details like the overall rate, and those extra fee.

2. We are just taking average price for each Airbnb, but we did not include the clean fee which could be real concern for many people while choosing the Airbnb

3.Overall rate: Best way to improve this prediction is to use the multiple regression line, but Data 8 does not include it.

Conclusion

In conclusion, we suspected that the distributions of the available days in Airbnb changed after COVID-19. Surprisingly, the distributions of the available days did not change after COVID-19 since the distribution is in our 95% confidence interval. If we had more data for years other than 2019 and 2020, we could see how the usual normal distribution is for each year, and get more accurate distribution on whether the change of distribution was big or not. We would have liked to perform analysis on how the availability is related to the price of the rent.

Reference

http://insideairbnb.com/get-the-data.html